INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 N4795
Fairfax, VA, USA, May 2002

Source: Requirements
Status: Draft
Editors: A. Smolic, R. Yamashita
Title: Draft Requirements for 3DAV

1 Introduction

The main goal of this activity is to provide certain new kinds of media that extend the functionalities of available standard technology. A key feature of these media is interactivity in the sense that the user shall have the possibility to chose his own viewpoint within a visual scene. Another feature covered by this activity is stereo vision that gives the user the impression of a 3D view of a visual scene. 3D Video can be defined as geometrically calibrated and temporally synchronized (group of) video data. Another definition might be image-based rendering using video input data or video-based rendering. This includes corresponding 3D Audio as well, which will also be considered in this activity.

So far 3DAV includes 5 main categories of scene representations:

1. Omni-directional (panoramic) video

This is an extension of the planar 2D image plane to a spherical or cylindrical image plane. Other kinds of planes (hyperbolic) are also possible. Video is captured at a certain viewpoint (which may move over time) into every direction. Any 2D view in any direction can be rendered from this representation. It can be applied to broadcast and storage (e.g. DVD) applications.

2. Interactive stereo video

2 views, one for each eye, are provided, to produce a 3D impression for the viewer. Head motion parallax can be supported to enable interactivity (in a certain operating range). It can be applied to broadcast, storage (e.g.

DVD), and communication applications. This can be considered as a special case of interactive multiple view video.

3. Interactive multiple view video

In this case a scene is captured by N cameras. Different camera settings are possible, e.g. parallel view, convergent view, divergent view, but in general any setting of cameras (e.g. combinations of the above) is possible, i.e. multiple view. Additional information (to the N video signals) about camera calibration and scene geometry (e.g. disparity data) enables interactive navigation through the scene. A simple case allows only to chose one of the predefined camera positions. In general one 2D view (for conventional displays) or 2 views (for stereoscopic displays) can be rendered from the data. This can be applied to broadcast and storage (e.g. DVD) applications, in simple cases also for communication applications.

4. 3D video objects

A scene is captured as in multiple view video (see 3), and one or more 3D video objects are created. A 3D video object comprises shape and appearance. The shape can be described by, e.g., polygon meshes, implicit surfaces, depth images, or multiple layered depth images. The appearance data is mapped onto the shape and allows the 3D video object to be seamlessly blended into new 2D or 3D video content. Appearance is typically described by a series of video streams, comprising textures, surface light fields (i.e., view-dependent textures), or surface reflectance fields (i.e., illumination- and view-dependent textures). The 3D video object can be composited into existing content, or it can be interactively viewed from different directions, or under different illumination. 3D objects can be applied to broadcast, storage (e.g., DVD), and interactive online applications.

5. 3D Audio

As the viewpoint of 3D video moves, the listening point and/or sound source position are/is also moves. 3D sound can be recorded by several ways.

First, multiple stereo microphones with cameras can make multiview sound. The multiview sound can be simply manipulated as movement of viewpoint, but this scheme needs huge memory for transmitting/storing all sound objects.

Second, 3D microphone can record all directional sound. In this case, it freely changes the listening point and sound source position, but it needs high computational load.

Third, object-based microphones can record 3D sound at their own positions by uni-directional microphones and the results can be coded individually. These sound objects need sound scene composition tool such as MPEG-4 advanced AudioBIFS to make a 3D sound scene. Each sound object needs 3D positioning tool such as HRTF rendering for mapping a monaural sound to 3D sound space. Thus it needs also 3D sound postion description tools.

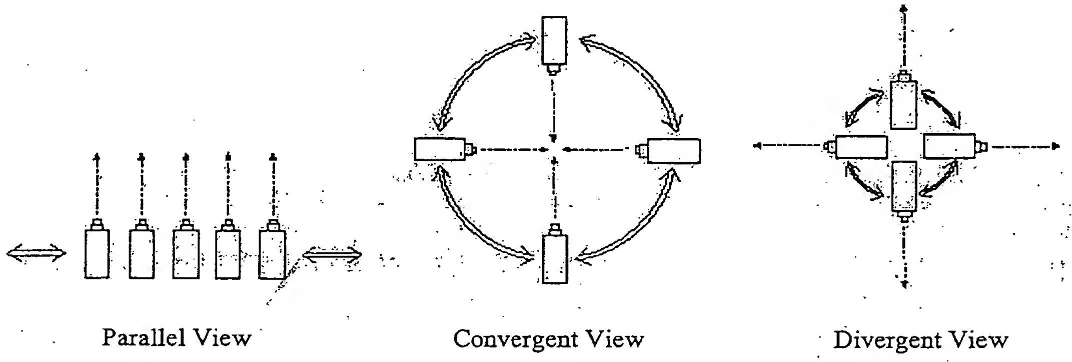

Fig.1. Types of spatial camera configuration for multiview video

Figure 2 shows a possible classification of 3D video, using acquisition, representation and display as criteria. The examples below show some concrete instantiations of this concept.

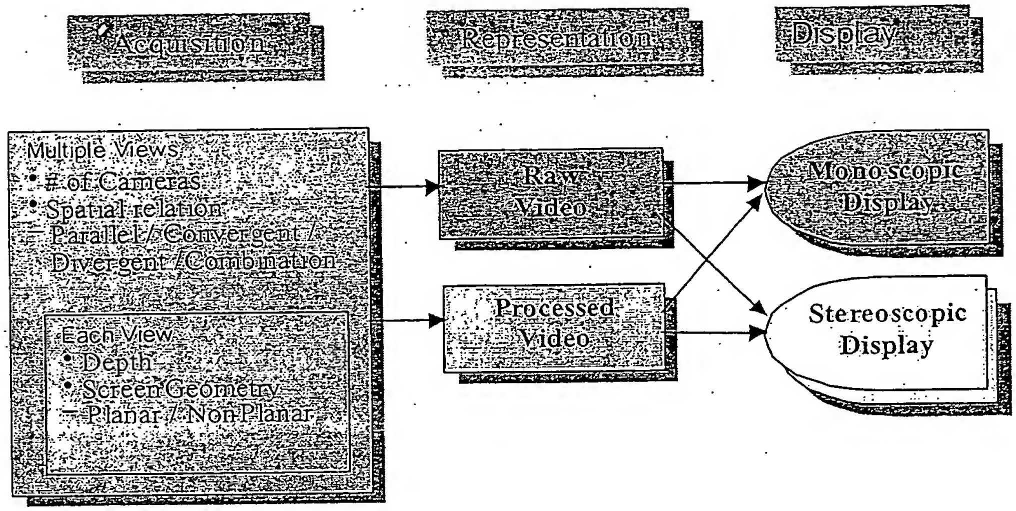

Fig.2. Classification of 3D video

Example1 Simple Stereoscopic Video

Acquisition	Representation	Display
Stereoscopic Video Camera • # of Cameras = 2 • Spatial relation: Parallel • Screen Geometry: Planar	Raw Video (2 Video Streams)	Stereoscopic Display

Example 2 Stereoscopic Video with Depth

Acquisition	Representation	Display
Video Camera with Depth • # of Cameras = 1 • Spatial relation: N.A. • Screen Geometry: Planar	Raw Video (1 video stream + Depth)	Stereoscopic Display

Example 3 Panorama View Video

Acquisition	Representation	Display
Omni View Video Camera • # of Cameras = 1 • Spatial relation: Divergent • Screen Geometry: Non-Planar	Raw Video (1 Video Streams)	Mono-Scopic Display

Example 4 Free Viewpoint Video (1)

Acquisition	Representation	Display
Video Camera • # of Cameras = 16 • Spatial relation: Convergent • Screen Geometry: Planar	Processed Video (Image Based Rendering Model)	Mono-Scopic Display

Free Viewpoint Video (2)

Acquisition	Representation	Display
2D Video Camera • # of Cameras = 13 • Spatial relation: Convergent • Screen Geometry: Planar	Processed Video (Light Field Model)	Mono-Scopic Display

Three different kinds of interactivity with the video can be distinguished:

1. Interaction at the encoder side

In this case, the end user selects the viewpoint by remotely interacting with the encoder side. Also, for the reduction of required channel bandwidth, the display type information can be used in display-dependent encoding. It can be applied to 1:1 AV delivery model such as video communications, VoD, webcasting, etc. In case of VoD and webcasting, in order to serve multiple users simultaneously, the sender should have multiple encoders in case of real-time encoding or have different compressed files depending upon viewpoints and display types in case of non-realtime encoding. The backchannel information can also be used in the multiplexing process to format proper bitstreams for video data required by users.

2. Interaction with all data available at the decoder side
In this case the end user has all video and additional data available and can navigate freely within the scene. This is practical for storage (e.g. DVD) applications and all kinds of interactive video. Practical solutions for broadcast applications and omni-directional (panoramic) and interactive stereo video are possible. Broadcast of interactive multiple view video might be impractical due to huge amount of data (increases with N).

3. Interaction without all data available at the decoder side
Here the end user does not have all video and additional data available. In this case, the video data of all viewpoints are compressed at the encoder side but only the bitstreams of video with user's requested viewpoint(s) and display types are transmitted to the decoder side. Hence, free navigation within the scene requires a backchannel which is impractical for broadcast applications. This is also out of scope for storage (e.g. DVD) applications. However, such an approach might be appropriate for streaming (e.g. Internet, Client/Server) applications to avoid initial download of all data.

1.1 3D-Video System Architecture

Fig.3 shows the block diagram of the 3D-Video system architecture. Coloured blocks indicate parts that have to be considered in the standardization process. The blocks have the following functionalities:

1. Video Capturing with Camera Parameters
Captures images and outputs raw video data and associated camera parameters, which might include depth data

2. Format Conversion
Converts raw data to uncompressed format

3. Data processing
Performs application specific processing functions and outputs such application specific data in the uncompressed format. Some examples of application specific processing are listed below:
   i) Integration with computer graphics
   ii) Construction of 3D models and texture data
   iii) Processing to support view dependent coding
   iv) Processing to support display dependent coding
   v) Extraction of depth, disparity data

4. Encoding
Transforms uncompressed format data into compressed format

5. Delivery
Delivery of compressed data via media such as broadcasting, network, DVD, etc.

6. 3D-Video Decoding
Decoding of the compressed data and output of uncompressed format data

7. Rendering
Renders the video on the screen

8. Interaction
Get the user's request to change the view point and view direction and transfer them to rendering part or 3D based processing part through backchannel.

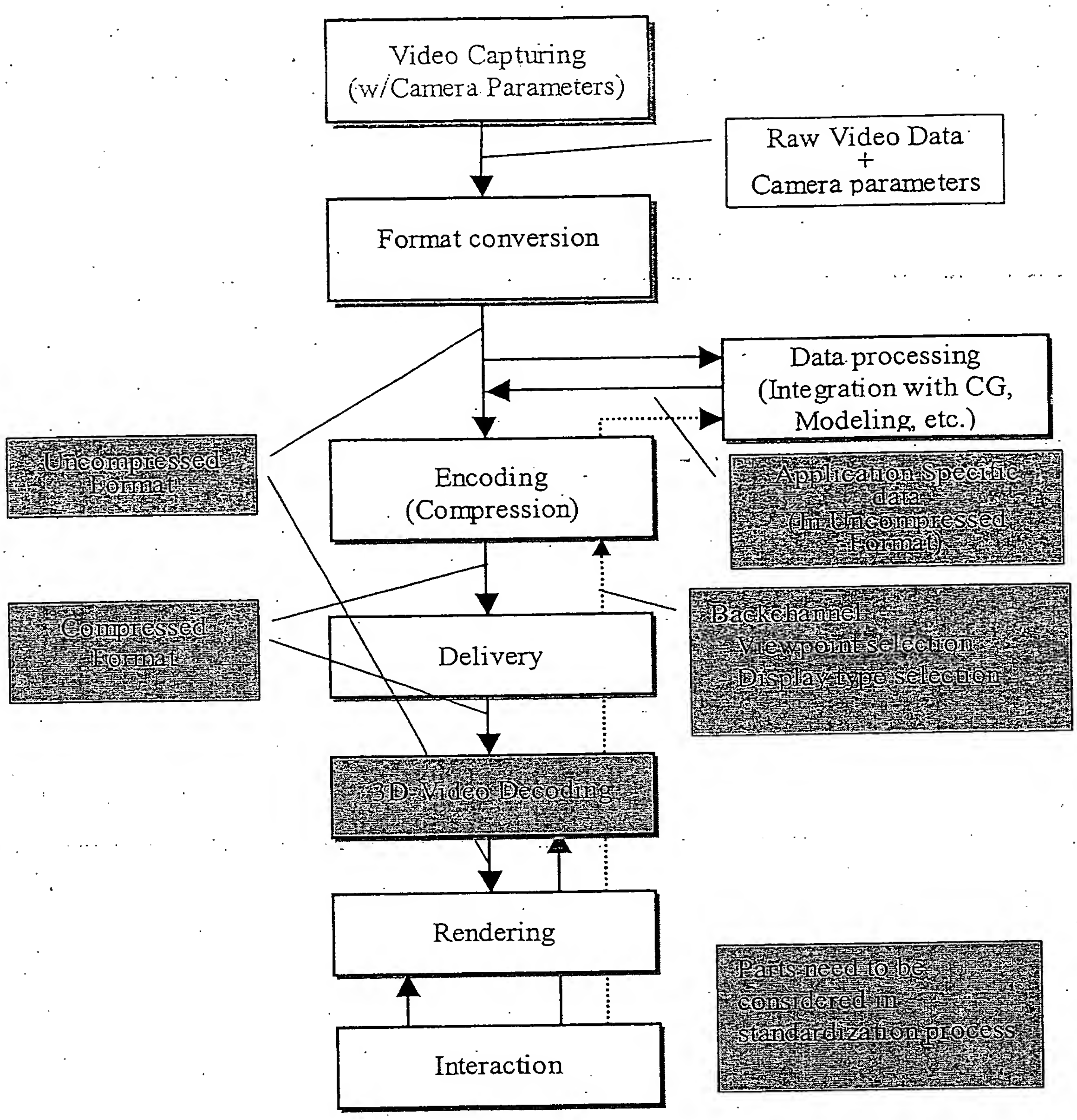

Fig.3. Block diagram of 3D video system architecture

1.2 Standardization items

The section gives a high-level overview of what needs to be standardized for the different categories of scene representations (without claim to be exhaustive, complete or consistent).

1. Omni-directional (panoramic) video
   - Panoramic video signal itself
   - Non-planar image plane (sphere, cylinder, other)
   - Mapping to 2D image plane

2. Interactive stereo video
   - (At least) one video signal
   - (3D) Shape and object information
   - Camera calibration information (might not be mandatory)
   - (Accurate) depth information (e.g. disparity data), possibly represented in new ways (might not be mandatory)

- Additional data, hidden layers (might not be mandatory)

3. Interactive multiple view video
- N video signals
- (3D) Shape and object information
- Camera calibration information (might not be mandatory)
- (Accurate) depth information (e.g. disparity data), possibly represented in new ways (might not be mandatory)
- Additional data, hidden layers (might not be mandatory)

4. Interactive 3D Audio
- N audio signals
- Natural 3D audio itself
- (3D) Microphone arrangement configuration (might not be mandatory)
- (3D) Audio scene description/composition
- Listening environment description
- Geometrical co-location of 3D audio and video
- Additional data (might not be mandatory)

1.3 Relation with available standards and on going activities

It has been identified that there are several tools already available in different MPEG standards or under investigation that are relevant for 3DAV. This section gives an overview of these tools and an analysis of their potential suitability for the 3DAV framework. However, this needs further investigation including technical evaluation and experimentation.

1.3.1 Overview

1. Omni-directional (panoramic) video

This could be handled by 3D BIFS by mapping of a video onto inner surface of a sphere or cylinder. However, conventional coding of panoramic video might not be efficient, special coding schemes taking into account spherical projection should be investigated. A special solution (without 3D BIFS) would be much more compact and easier (cheaper) to handle/implement.

2. Interactive stereoscopic video

A multi-view profile is available in MPEG-2 but it does not support interactivity at all. The Multiple Auxiliary Components (MAC) from MPEG-4 can be used for disparity data, representation with more than 8 bit and loss-less coding are possible. But MAC might not be efficient due to special nature of disparity data. New methods for representation and coding of depth data should be investigated, e.g. layered depth images. Some of this technology is already under investigation in SNHC. Certain kinds of camera calibration information are available in MPEG-4 and MPEG-7, however these do not satisfy the needs of interactive video in terms of functionality and accuracy. New descriptors in MPEG-7, Sensor Parameters and 3D Coordinates, might be appropriate for interactive video, however these are under investigation in CE status, and the CE might be stopped due to lack of activity. Hidden layers could be coded as arbitrary shaped MPEG-4 VOPs, however due to the nature of the data (irregular shaped, many small pieces) this might not be efficient, new approaches should be investigated. This means that some useful tools are already available or under investigation in MPEG. Some new should be developed to increase efficiency. The elements are spread over different standards and groups. They need to be combined in a special framework for interactive stereo video.

3. Interactive multiple view video

In addition to what is stated for interactive stereo video a framework needs to be developed to accommodate the scene description with N calibrated cameras.

4. 3D Audio

MPEG-4 advanced AudioBIFS can make interactive 3D sound scene. Sound and DirectiveSound node spatialize a sound object to 3D sound space.

1.3.2 Depth image-based representation (AFX)

The MPEG-4 Animation Framework eXtension (AFX) defines two image-based depth representations for use in image-based rendering. A simple Depth Image (DI) which is just an image with an associated depth map, as well as a multivalued image and depth representation, called a Layered Depth Image (LDI). A simple DI allows to create novel views of a scene by means of 3D warping of the DI pixels. In addition to this, an LDI allows to store additional depth and color values for pixel that are occluded in the original view. This extra data provides the necessary information that is needed to fill disoccluded areas in the rendered, novel views.

Up to now AFX specifies a DepthImage structure, which consists of a computer-graphics centric camera definition (i.e. position, orientation, field of view, near clipping plane, far clipping plane, etc.) and a pointer to a depth image. This can either be a SimpleTexture (i.e. a DI) or a PointTexture (i.e. an LDI). For a SimpleTexture, the texture and depth fields can be comprised of either an ImageTexture, a MovieTexture or a PixelTexture, as defined in the MPEG-4 Video/System documents. A PointTexture is comprised of a) a texture which stores for each pixel the number of layers as well as the color values for each layer; b) a depth map which stores for each pixel the 4-byte depth values for each layer. In either case, the depth values should be normalised to the distance from the near to the far clipping plane of the camera.

The authors explicitly note that the format could also be used for animated objects (i.e. sequences) by storing sets of compressed video streams instead of images, together with 'streams' of depth maps. For compression, they simply state that still and video coding formats of MPEG-4 should be used for textures and depth maps.

Source:

MPEG-4 Animation Framework eXtension (AFX) VM 6.0 (N4626)

Status:

Under consideration

Suggested reading:

L. McMillan. An Image-Based Approach to Three-Dimensional Computer Graphics. PhD thesis, University of North Carolina at Chapel Hill, 1997.

M. Oliveira, G. Bishop, and D. McAllister. Relief Texture Mapping. In Proceedings of SIGGRAPH '00, pages 259-268, New Orleans, LA, USA, July 2000.

J. Shade, S. Gortler, L.-W. He, and R. Szeliski. Layered Depth Images. In Proceedings of SIGGRAPH '98, Orlando, FL, USA, July 1998.

N. L. Chang and A. Zakhor. Constructing a Multivalued Representation for View Synthesis. International Journal of Computer Vision, 45(2):157-190, 2001.

C.-F. Chang, G. Bishop, and A. Lastra. LDI Tree: A Hierachical Representation for Image-Based Rendering. In Proceedings of SIGGRAPH '99, Los Angeles, CA, USA, August 1999.

Assessment:

The depth image-based representation, as currently under consideration within the MPEG-4 Animation Framework eXtension (AFX), seems very much computer graphics oriented (see e.g. the camera definitions) at the current stage. Nevertheless it seems worthwhile to have a closer look at it, as both the simple Depth Image (DI) and the Layered Depth Image (LDI) seem to fit very well into the 3DAV framework. A joint discussion between the 3DAV and the AFX groups seems necessary and useful, as the proposed techniques still seem to be at a very early stage of development. Especially the AFX authors statement that 'for compression still and video coding formats of MPEG-4 should be used for textures and depth maps', seems rather vague and certainly needs more investigation and discussion.

1.3.3 Multiview profile (MPEG-2)

The MPEG-2 multiview profile (MVP) was defined in 1996 as an amendment to the MPEG-2 standard with the main application area being stereoscopic TV. The MVP extends the well-known hybrid coding towards exploitation of inter-viewchannel redundancies by implicitly defining disparity-compensated prediction. The main new elements are the definition of usage of the temporal scalability (TS) mode for multi-camera sequences, and the definition of acquisition parameters in the MPEG-2 syntax. The TS mode was originally developed to allow for the joint encoding of a low frame rate baselayer stream and an enhancement layer stream comprised of additional video frames. If both streams are available an decoded, video could be reproduced with full frame rate. In the TS mode, temporal prediction of enhancement layer macroblocks could be performed either from a base layer frame, or from the recently-reconstructed enhancement layer frame.

For stereo or multichannel signals comprised of the video data captured simultaneously from two or more views of the scene, it is straighforward to perform encoding using the TS syntax. For this purpose, frames from one camera view are defined as the base layer, and frames from the other one(s) as enhancement layer(s). The enhancement-from-base-layer prediction then turns out as a disparity-compensated prediction instead of a motion-compensated prediction. If the disparity-compensated prediction fails, it is still possible to achieve compression by motion-compensated prediction within the same channel. At the same time, the base layer represents a monoscopic sequence.

Unfortunatly, disparity vectors defined on a block-by-block basis of 16x16 pixels, as used in the TS mode of MPEG-2, are not accurate enough to minimize the inter-viewchannel prediction error to the possible extent. It can be observed that in many cases (with exception of high motion) the similarity between subsequent frames within one of the views is much higher than the similarity between the different views, such that the motion-compensated interframe prediction is most likely preferred over the disparity-compensated inter-viewchannel prediction. As a consequence, the temporal scalability concept can only be marginally superior over a separate encoding (so-called simulcast) of the channels, both concepts requiring approximately doubled rate as compared to encoding a signal from a single camera.

Source:
Final Text of 13818-2/AMD 3 (MPEG-2 Multiview profile) (N1366)

Status:
Already standardized

Suggested reading:
J.-R. Ohm. Stereo/Multiview Encoding Using the MPEG Family of Standards. Invited Paper, In Proceedings of Electronic Imaging '99, San Diego, USA, January 1999.

Assessment:
The MPEG-2 multiview profile (MVP) was developed as a tool for the coding of stereoscopic and multiview video sequences. As such it exploits inter-viewchannel redundancies by implicitly defining disparity-compensated prediction. MVP was not developed for the coding of depth maps or disparity information and therefore doesn't seem to be very useful for the current 3DAV activities where it is intended to transmit the basic video data together with associated depth or disparity information. The MPEG2 multiview profile is not 3D video because no geometric calibration data is associated with the video or hence no compression based on the 3D geometric redundancy is employed. It is also not possible to realize any viewpoint modification, i.e. interactivity is not supported at all.

In the MPEG-2 MVP, basically temporal scalability is used and thus each view is carried in each layer. In this case, synchronization between the views is achieved using timestamps. The multi-layer approach causes little problems in case of hardware decoders. However, view-synchronization based on the timestamp mechanism is very difficult to implement in case of software players running on PC platforms.

1.3.4 Multiple auxiliary components (MPEG-4)

The basic idea of the Multiple Auxiliary Components (MAC) is that grayscale shape is not only used to describe the transparency of the video object, but can be defined in a more general way. MACs are defined for a video object plane (VOP) on a pixel-by-pixel basis, and contain data related to the video object, such as disparity, depth, and additional texture. Up to three auxiliary components (including the grayscale or alpha shape) are possible. Only a limited number of types and combinations are defined and identified by a 4-bit integer so far, but more applications are possible by selection of a USER DEFINED type or by definition of new types. All the auxiliary components can be encoded by the shape coding tools, i.e. the binary shape coding tool and the gray scale shape coding tool which employs a motion-compensated DCT, and usually have the same shape and resolution as the texture of the video object.

Source:
Text of ISO/IEC 14496-2 (MPEG-4 Visual) 2001 Edition (N4350)

Status:
Already standardized

Suggested reading:
J.-R. Ohm. Stereo/Multiview Encoding Using the MPEG Family of Standards. Invited Paper, In Proceedings of Electronic Imaging '99, San Diego, USA, January 1999.

J.-R. Ohm, and K. Müller. Core Experiments on Multiview Objects. Doc. number M3178, February 1998.

R. Koenen. MPEG-4 Overview. Doc. number N4030, March 2001.

Assessment:

The MPEG-4 multiple auxiliary components (MAC) is a generalization of the grayscale shape coding. As such there are a number of possible shortcomings when used to encode depth maps or disparity information. First, the usage of MAC implies that the binary shape of the video object has to be transmitted. If there is none, the shape of the rectangular screen has to be encoded which is quite a waste of bits. Second, depth maps or disparity information have very specific characteristics and it is questionable if the artifacts introduced through DCT encoding and quantization are tolerable with respect to the quality of the synthesized views. Third, while it seems useful to use the texture motion vectors to compensate the depth or disparity sequence, so that one doesn't have to transmit a whole new lot of motion vectors for the auxiliary sequence, it is questionable if texture motion fields and depth or disparity motion fields are indeed always as close as one might think (e.g. think about the cases where texture changes due to illumination changes, while the scene geometry and thus the depth or disparity basically stays the same). If this is not the case, using texture motion vectors to predict depth or disparity images might lead to a very costly differential coding.

Despite these possible shortcomings, it seems useful to have a closer look at the MAC to see if they could be extended to better fit the needs of the 3DAV activities.

While MAC can be used to synthesize such object images that can be observed from intermediate viewpoints between the cameras and integrate them with 3D graphics data, it is not a full 3D representation, because no explicit 3D geometric calibration data is associated. A disparity map is incomplete 3D information. It is questionable if such a method would provide a sufficient compression efficiency.

1.3.5 Light Field Mapping (MPEG-4)

At Pattaya meeting, R.Grzeszcuk et al (M7604) proposed Surface Light Field Mapping. It defines a coding method for multiview object images associated with precise geometric calibration data. Instead of incomplete 3D information such as disparity, it includes an explicit 3D object shape represented by triangular patches, based on which textures on the patches can be generated depending on interactively specified viewpoints. Such explicit 3D representation enables very high data compression rate (100 – 3000 : 1). That is, while a large amount of image data are captured, they have significant 3D spatial redundancy in addition to 2D surface (texture) redundancy: each point on the object surface is recorded many times in multiview images.

While the surface light field mapping method is effective for STATIC 3D objects, it cannot be applied to moving objects, because the object shape and pose change dynamically.

1.3.6 Other tools to be investigated

The following tools need further investigation regarding their relevance and suitability for 3DAV:

- 3D-BIFS for video and audio
- Camera geometry information based on SMTPE315M
- Sensor Parameters and 3D Coordinates (proposed descriptors)
- MPEG-7 Camera Motion descriptor

2 Applications and products

2.1 Available today

The following list shows examples of applications available today on the market. Some of them are not directly related to 3DAV, however, they provide some sense of what 3D-AV is trying to realize.

1. Camera(+System)

   Forth View (Sony) – System for PS2
   http://www.sony.co.jp/Products/fourthview/

   Zcam (3DV systems) – Camera with Depth Data
   http://www.3dvsystems.com/

   DigiclopsTM (Point Grey Research)

http://ptgrey.com

FLYCAM(Fuji Xerox)
http://www.ubiquitous-media.com

Panorama Video Cam (Sharp)
http://www.sharp.co.jp/corporate/news/010919-2.html

EYE Vision (CBS, Mitsubishi Heavy Industry)
http://www.sdia.or.jp/mhikobe/products/video4d/indexj.html

CAM-4000 3D Camera (VREX)
http://www.vrex.com/products/cm_4000.shtml

2. Display

BOOM 3C (Fakespace)
http://www.fakespacelabs.com/products/boom3c.html

Sanyo 3D Screen (Sanyo)
http://www.sanyo.co.jp/koho/hypertext4/0109news-j/0912-1.html

3D-LCD (Philips)
http://www.research.philips.com/generalinfo/special/3dlcd/

3D TFT-LCD (Samsung)
http://www.sec.samsung.com/news/digital_media/

Cyberbook, Stereo3D-notebook (VREX)
http://www.vrex.com/products/micropol.shtml

21 MX (Nuvision)
http://www.nuvision3d.com/shutters.html

3. 3D glasses

StereoGraphics Corporation
http://www.stereographics.com/

Iart3d
http://www.iart3d.com/

Anotherworld
http://www.anotherworld.to/

i-O Display Systems
http://www.i-glasses.com/

VRJoy
http://www.vrstandard.com/

4. Panorama view software

Quicktime VR(Apple Computer) – divergent/convergent view still picture
http://www.apple.com

Be Here Technologies – divergent view movie
http://www.behere.com/

IPIX (Internet Pictures Corporation) – divergent view still picture http://www.ipix.com/

2.2 Possible applications in the near future

2.2.1 Multiple-view video

Interactively changing viewpoint and view direction in 3D video data. Any existing 2D video application can be replaced with 3D.

For very instance, viewing interactively changing viewpoint and view direction from cameras on players in football games, or the same from cameras on cars or drivers in F1 circuit. This may be more interesting in live broadcast, and more at multi-user communication environment.

Note: with surrounding display (3D display), interactivity is not necessarily mandatory. (But anyway this is one case subject to interactivity).

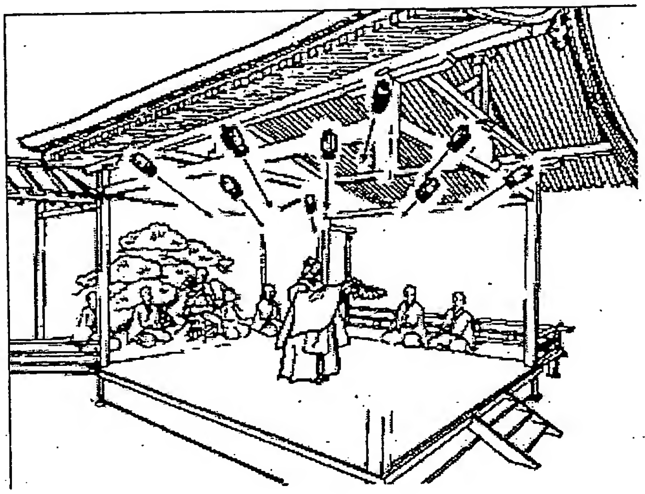

Fig.4. Recording of Noh Play with multiple views

Possible application domains are:

1. Entertainment
   - Concert
   - Sport
   - Disco
   - Multi-user Game
   - Movie

2. Education
   - Cultural Archives
   - Manual with real video
   - Instruction of sports playing

3. Medical surgery

4. Viewing with Exploration
   - Zoo, Aquarium, Botanical garden
   - Museum
   - Catalogue with real video (3D TV shopping)

5. Communication

6. Sightseeing

7. Surveillance

2.2.2 Stereoscopic (interactive) video

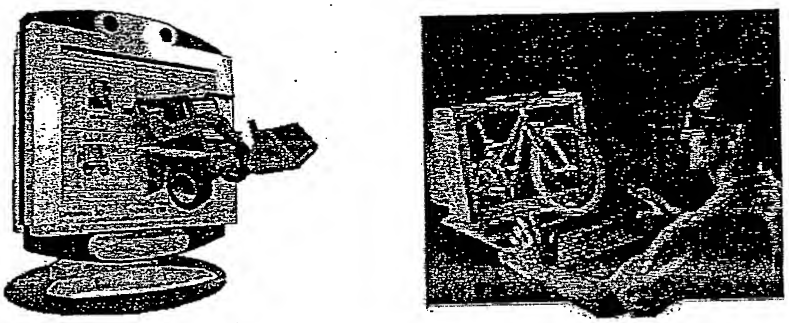

Fig.5. Examples of stereoscopic video

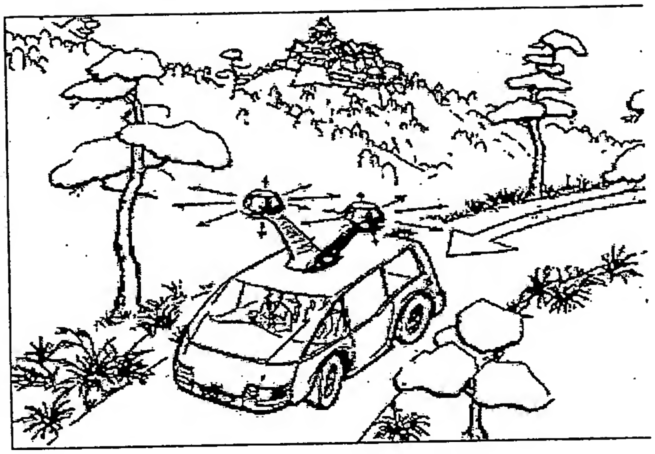

Fig.6. Recording scenery with stereo omni-direction video

1. Sports broadcast or webcasting
   - There were already the exhibition broadcast in the Olympic games of Sydney and Nagano.
   - ETRI has a plan to broadcast the games in 2002 World cup of Korea-Japan, experimentally.
   - In the case of soccer game, we can have a feeling as if we were really in the stadium.

2. Education
   - Driving, flight simulation, anatomy, molecular structure
   - Provide the reality in remote education.
   - Provide the cost effectiveness in some areas that need expensive material and apparatus.
   - In the case of driving education, beginner can practice without feeling the risk of accident.
   - In the case of molecular structure study, it is easier to recognize perspective relations among molecules.

3. Entertainment
   - Movie, game, home shopping, sight seeing
   - There are many stereoscopic theaters in the amusement park
   - There are many game products supporting stereoscopic functionality in the commercial market.
   - we can enjoy more realistic impression in movie, game, sight seeing, etc.
   - In the case of home shopping, stereoscopic impression gives consumer more visual information of products.

2.2.3 Broadcast 3D video

Probably 3D will be the next major revolution in the history of TV. Both at professional and consumer electronics exhibitions, companies are eager to show their new 3D products that always attract a lot of interest. Obviously, if a workable and commercially acceptable solution can be found, the introduction of 3D-TV will generate a huge replacement market for the current 2D-TV sets. In this decade, it can be expect that technology will have progressed far enough to make a full 3D-TV application available on the mass consumer market, including content generation, coding, transmission and display.

The ATTEST consortium works towards a flexible, 2D-compatible and commercially feasible 3D-TV system for broadcast environments. The consortium consists of 8 European partners (e.g. Philips and HHI).

As the main depth cue comes from stereovision, it seems natural to record and distribute 3D broadcast signals as separate video streams for each eye. Hence existing trials for the introduction of 3D-TV are based on this idea of 'stereoscopic' video. This approach is merely broadcast production centered and restricted in some sense: the director has full control over the depth effect; the viewer has to accept the effect as provided. Backwards compatibility is not necessarily supported and there is no possibility to change the viewing conditions, to scale the grade of depth perception or to adapt to different kinds of mono, stereo or multi-view displays.

In contrast, to cope with the different aspects of compatibility, scalability and adaptability, the ATTEST approach to 3D-TV is quite different from former ones. The core is a flexible and scalable syntax for image-based 3D data representation, which will be open for different display types and viewing conditions (see Figure 7). The purpose of this syntax is to introduce 3D video as a combination of regular video (2D base layer) and synchronized image-based depth information (3D enhancement layers). Due to this structure our 3D-TV approach will be backward compatible to existing 2D video services.

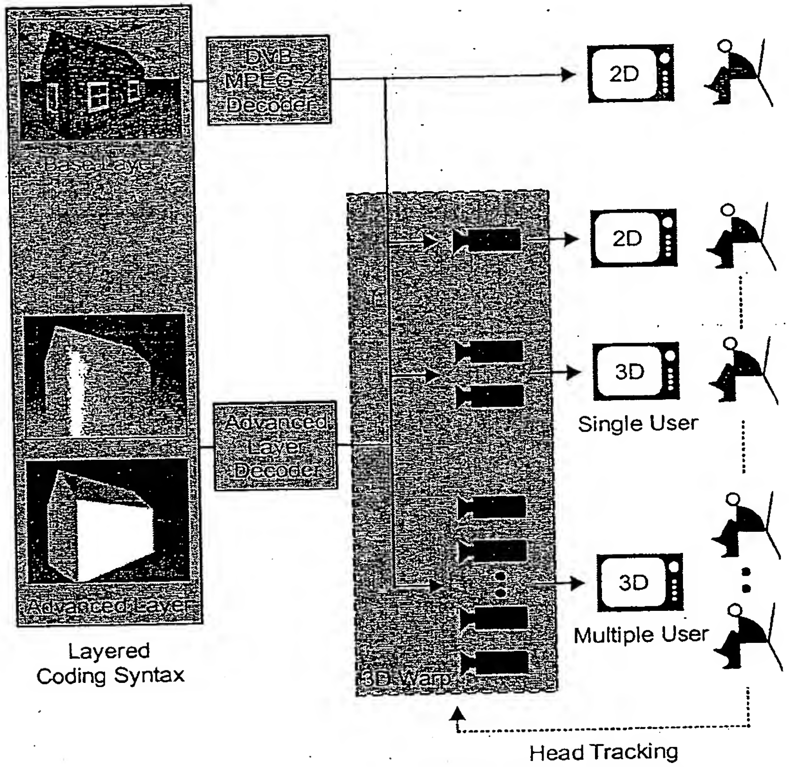

Fig.7. The layered coding syntax provides backward compatibility to conventional 2D digital TV and allows to adapt the view synthesis to a wide range of different 2D and 3D displays.

It will further be possible to reconstruct and to control multiple virtual views, supporting a wide variety of tracked, mono, stereo and multi-view displays. Hence, the system will be scalable in terms of receiver complexity - an important issue to introduce 3D-TV in an evolutionary manner. Due to the usage of 3D enhancement layers, the syntax will also provide scalability in terms of depth experience. This is particularly important because perception studies have indicated that there are differences in depth appreciation over age groups. Hence in our view, the viewer should be in control of his depth experience. He should be able to set the depth level according to his personal preference - a feature that can also be used as graceful degradation in the case of unexpected artifacts in depth which are usually more annoying in stereovision than in parallax viewing.

3 3DAV requirements

Note: The requirements need to be clustered in a meaningful way, e.g. video, audio, systems, etc.

1. **Uncompressed representation format of 3D video**
3DAV shall define an uncompressed format for representation of 3D video content.

2. **Compressed format of 3D video**
3DAV shall provide a compressed format for exchanging 3D video content between different systems.

3. **Uncompressed representation format of 3D audio**
3DAV shall define an uncompressed format for representation of 3D audio content.

4. **Compressed format of 3D audio**
3DAV shall provide a compressed format for exchanging 3D audio content between different systems.

5. **Extrinsic and intrinsic camera parameters**
Camera parameters shall be described which contribute to the reconstruction of 3D views. This shall include extrinsic parameters such as 3D position and angle as well as intrinsic parameters (focus, aspect ratio, etc.). This shall enable full geometric calibration of the imaging system in 3D. Dynamic changes shall also be represented (e.g. camera rotation and translation, zoom). Those parameters may include:

- Camera Models
  - Geometric Information
  - Photometric Information
  - Temporal Information
  - Screen definitions
- Camera Types
  - Pin-hole camera
  - Thin lens camera
  - Thick lens camera
  - Fish-eye lens camera
  - Camera with lens and mirror(s)
- Camera Works
  - Motion (rotation, transition, position)
  - Zooming
  - Focusing
  - Iris and gain control

6. **Non-planar imaging and display systems**
3DAV shall support efficient representation and coding methods for non-planar imaging and display systems. This shall include for instance cylindrical spherical image planes. These methods should be designed taking into account that the video can be projected easily and efficiently onto non-planar screens.

7. **Multiple Views**
Multiple views of a scene shall be described. This shall include stereoscopic views, i.e. views with associated depth or disparity data.

8. **Synchronization**
Accurate temporal synchronization between the multiple views shall be supported.

9. **Reuse of existing tools**
Wherever possible existing MPEG tools shall be reused.

10. **Integration with SNHC computer graphics objects**
The 3DAV framework shall allow integration of existing SNHC computer graphics objects.

11. **Interactivity**
Interactive change of viewpoint and view angle shall be supported. This shall include local interactivity at the decoder and remote interactivity between decoder and encoder. The latter requires a backchannel.

12. **Disparity, depth information**
Efficient representation and coding of depth maps (e.g. disparity data) should be supported, enabling loss-less reconstruction, highly accurate depth representation and efficient compression at the same time. Such data are used to reconstruct a stereo-pair VOP.

13. **3D objects**
3D objects shall be supported for handling several stereoscopic or multiview objects in a scene.

14. **Backwards compatibility**
Backwards compatibility with MPEG-2 video should be supported, since 2D and 3D broadcast will co-exist while introducing 3D-TV.

15. **Compression efficiency**
3DAV shall provide high compression efficiency for a wide range of applications. This shall include broadcast as well as mobile communication scenarios. The overhead by additional 3D data should be limited (to e.g. 20%), in order to increase acceptance of the new services.

16. **Performance efficiency**
3DAV shall be efficient in terms of computational complexity.

17. **Occlusion handling**
Efficient representation and coding of multi-label masks and hidden layers for occlusion handling in interactive stereo video should be supported.

18. **Different display types**
Different 3D displays should be supported. This shall include conventional 2D displays, field- and frame-based shuttering displays, conventional stereo vision (non-tracked, e.g. using glasses), head motion parallax viewing on 2D displays, single-user viewing including head-motion parallax (head-mounted displays, auto-stereoscopic displays), multi-user auto-stereoscopic viewing on large screens. This shall also enable the user to change the display interactively.

19. **Scalability**
Complexity (i.e. cost) scalability of the end user terminals shall be supported.